

Agnieszka Zięba

Szkoła Główna Handlowa w Warszawie

WYBRANE NIEKLASYCZNE METODY ESTYMACJI: MONTE CARLO, BOOTSTRAP, JACKKNIFE

1. Wstęp

Standardowe metody wnioskowania statystycznego zakładają losowy dobór próby, podczas gdy większość badań społecznych i marketingowych opiera się na próbach kompleksowych. W wielu przypadkach, mimo iż próba losowa jest niemożliwa do zrealizowania, wnioskowanie przebiega przy założeniach takich jak dla próby losowej. Efektem takiego postępowania może być uzyskanie obciążonych wyników badania [4]. Poprawne wnioskowanie na podstawie prób kompleksowych umożliwiają m.in. tzw. nieklasyczne metody wnioskowania, a zwłaszcza metody bootstrap i jackknife będące adaptacją metody Monte Carlo do praktycznych rozwiązań zagadnień statystycznych [5].

Metody bootstrap i jackknife wykorzystywane są również w zagadnieniach imputacji danych. W przypadku prób kompleksowych potrzeba imputacji pojawia się m.in. w wyselekcjonowanych wcześniej podgrupach, w których ze względu na brak znajomości rozkładu nie możemy stosować metod klasycznych.

Celem pracy jest porównanie trzech nieklasycznych metod wnioskowania statystycznego: Monte Carlo, jackknife i bootstrap. Od metod klasycznych różnią się one wykorzystaniem danych, choć oparte są na tej samej idei. Niniejsza praca zawiera zestawienie algorytmów działania tych metod oraz sposobów wykorzystania metod jackknife i bootstrap do analizy danych niepełnych.

Metoda estymacji Monte Carlo jest metodą numerycznego rozwiązywania zadań, dostarczającą przybliżonych rozwiązań problemów matematycznych dzięki przeprowadzanym komputerowo eksperymentom na próbach statystycznych. Cechą dla niej charakterystyczną jest badanie zjawiska przez jego obserwacje dla wielu powtórzeń eksperymentu losowego, co stanowi punkt wyjścia dla metod jackknife i bootstrap [5].

Szerokie obszary zastosowań metod jackknife i bootstrap potraktowane zostaną wybiórczo. Przedstawione tu problemy będą związane z zagadnieniami imputacji danych. W niniejszych rozważaniach dla estymatora w postaci funkcji elementów próby skupimy się na szacowaniu obciążenia oraz wariancji służących badaniu jakości estymatorów. Charakterystyki te w obliczu prezentowanych metod umożliwiają m.in. określenie precyzji estymatora oraz porównanie danego estymatora z innymi estymatorami tej samej cechy populacji. Rozważania prowadzone będą dla funkcji elementów prostej próby losowej $X_n = (X_1, \dots, X_n)$, czyli statystyki $\hat{\theta}_n = T(X_n)$. Ciąg wartości $x_1, x_2, \dots, x_n$ oznaczać będzie konkretną realizację prostej próby losowej.

2. Metody Monte Carlo

Techniki stosowane w metodach Monte Carlo mają bardzo długą historię. Te używane współcześnie powstały jednak w okresie drugiej wojny światowej w USA przy okazji rozwiązywania zadań związanych z fizyką atomową. Nazwę i systematyczny rozwój tych metod datuje się od 1944 r., kiedy to J. von Neuman i S. Ulam prowadzili badania dotyczące bomby atomowej [3].

Metody Monte Carlo mogą być wykorzystane do estymacji w sytuacji, kiedy znany jest nam rozkład prawdopodobieństwa opisujący interesujące nas zjawisko. Rozkład ten zadany przez dystrybuantę F umożliwia generowanie prostych prób losowych.

Jeżeli interesuje nas estymator $\hat{\theta}_n = T(X_n)$ parametru θ i chcemy zbadać jego jakość w pierwszym etapie z rozkładu o dystrybuancie F (za pomocą generatora liczb pseudolosowych) generujemy k prób losowych o liczebności n każda. Następnie dla każdej z prób oddzielnie obliczamy wartość estymatora $\hat{\theta}_n$. Wówczas histogram wartości $\hat{\theta}_n^1, \hat{\theta}_n^2, \dots, \hat{\theta}_n^k$ jest dla dużej liczby k dobrym przybliżeniem rozkładu statystyki $\hat{\theta}_n$. Estymator wariancji dla $\hat{\theta}_n$ przybiera postać:

$$\hat{v}_{MC} = \frac{1}{k(k-1)} \sum_{i=1}^k (\hat{\theta}_n^i - \hat{\theta}_{(\cdot)})^2, \quad \text{gdzie} \quad \hat{\theta}_{(\cdot)} = \frac{1}{k} \sum_{i=1}^k \hat{\theta}_n^i.$$

Jednak z reguły nie znamy dystrybuanty F i dysponujemy jedynie pojedynczą realizacją próby tego rozkładu. W sytuacji, gdy nie jesteśmy w stanie powiedzieć nic o postaci tego rozkładu, zastosowanie znajdują metody jackknife i bootstrap [5].

Zarówno jackknife, jak i bootstrap są metodami symulacyjnymi. Ze względu na powtarzalność eksperymentu są one adaptacją metody Monte Carlo w przypadku, gdy dysponujemy jedną realizacją próby i nie wiemy nic o postaci rozkładu, z którego próba ta pochodzi. Jackknife i bootstrap znajdują zastosowanie m.in. w estymacji rozkładów statystyk oraz ich charakterystyk. Ich cechą charakterystyczną jest powtórne próbkowanie, czyli tworzenie prób nowych na podstawie

próby danej. W związku z tym zaliczane są one do metod tzw. resamplingowych. Ich zaletą jest to, że przy minimalnej liczbie założeń dają dokładne wyniki – estymatory otrzymane przy ich zastosowaniu są zgodne [2].

3. Metoda jackknife

Metoda jackknife została przedstawiona w roku 1949 przez Maurice’a Que-nouille’a [7], który pierwszy raz użył jej do estymacji obciążenia estymatora. Znaczenie tej metody wzrosło, gdy w roku 1958 twórca jej nazwy J.W. Tukey odkrył, że można jej użyć do estymacji wariancji estymatora. Jackknife w języku potocznym oznacza kieszonkowy nóż. Tukey skojarzył go z odcinaniem kolejno po jednej obserwacji z danej próby. Właśnie w ten sposób bowiem konstruuje się próbę charakterystyczną tylko dla metody jackknife.

W celu uzyskania próby typu jackknife $X_{n-1,i}$ spośród elementów oryginalnej próby $X_n = (X_1, \dots, X_n)$ usuwamy element X_i . Czynność tę powtarzamy n razy, w wyniku czego otrzymujemy n takich prób, z których każda przybiera postać $X_{n-1,i} = (X_1, \dots, X_{i-1}, X_{i+1}, \dots, X_n)$.

Zagadnieniem, na jakim się teraz skupimy będzie estymacja zarówno parametru θ , jak i jego obciążenia oraz wariancji (szczegółowe informacje opisujące poniższy tok postępowania znajdują się w pozycjach [2; 9]). Zakładamy, że potrafimy obliczyć wielkość estymatora $\hat{\theta}_n = T(X_n)$. Chcemy poznać obciążenie i błąd standardowy tego estymatora. Krokiem zbliżającym nas do tego jest obliczenie replik typu jackknife postaci $\hat{\theta}_{n-1,i} = T(X_{n-1,i})$, czyli wartości estymatora dla każdej z prób typu jackknife, gdzie $X_{n-1,i}$ jest próbą opartą na $n-1$ elementach próby X_n . Histogram utworzony za pomocą wartości $\hat{\theta}_{n-1,i}$ jest przybliżeniem rozkładu estymatora $\hat{\theta}_n$, co umożliwi nam skonstruowanie w sposób opisany poniżej estymatorów obciążenia i wariancji.

Estymator obciążenia typu jackknife dla $\hat{\theta}_n$ przybiera postać:

$$bias_{jack} = (n-1)(\hat{\theta}_{(\bullet)} - \hat{\theta}_n), \quad \text{gdzie} \quad \hat{\theta}_{(\bullet)} = \frac{1}{n} \sum_{i=1}^n \hat{\theta}_{n-1,i}.$$

Wtedy estymator parametru θ o zredukowanym obciążeniu to:

$$\hat{\theta}_{jack} = n\hat{\theta}_n - (n-1)\hat{\theta}_{(\bullet)}.$$

Estymatorem wariancji typu jackknife dla $\hat{\theta}_n$ nazywamy estymator postaci:

$$\hat{v}_{jack} = \frac{n-1}{n} \sum_{i=1}^n (\hat{\theta}_{n-1,i} - \hat{\theta}_{(\bullet)})^2.$$

Warunkiem tego, aby estymator mógł być użyty, jest jego zgodność. Zgodność ta jest dowodzona dla estymatorów w postaci funkcji średniej $\bar{X}_n = \frac{1}{n} \sum_{i=1}^n X_i$, czyli $\hat{\theta}_n = g(\bar{X}_n)$, w których g spełnia kryteria gładkości funkcji. Szczegóły dotyczące tego dowodu zawarte są w pozycji [9]. Działania te podporządkowane są ogólnej technice dowodzenia zgodności estymatorów i mogą być wykorzystywane do dowodu zgodności także dla innych niż średnia estymatorów postaci $\hat{\theta}_n = g(\bar{X}_n)$.

3.1. Metoda jackknife w analizie danych niepełnych

Wykorzystanie metody jackknife w analizie danych niepełnych wiąże się z uprzednim wyborem jednej z metod imputacji danych. Chcąc obliczyć wartość oraz wariancję interesującej nas statystyki, metodę imputacji należy zastosować na poziomie tworzenia prób typu jackknife. Budując każdą z takich prób, najpierw usuwamy element i -ty próby, czyli X_i . Każda z tak utworzonych prób zawiera braki. Braki te uzupełniamy wybraną metodą imputacji (dla każdej próby typu jackknife $X_{n-1,i}$ oddzielnie). Możemy zastosować tu np. metodę imputacji średniej ogólnej lub też losową metodę „hot deck”. Na podstawie tak utworzonych prób konstruujemy kolejno repliki $\hat{\theta}_{n-1,i}$, estymator statystyki $\hat{\theta}_{(n)}$ oraz estymator wariancji dla $\hat{\theta}_n$ [9].

Jeżeli metodą zastępowania braków danych będzie losowa metoda imputacji „hot deck” estymator zgodny wariancji otrzymamy wówczas, gdy zastosujemy następujące podejście. Losowa metoda „hot deck” polega na tym, że miejsca braków danych wypełniane są przez elementy prostej próby losowej wylosowanej spośród zbioru wszystkich wartości zaobserwowanych. Aby jednak metoda ta dawała w efekcie zgodne estymatory, należy wprowadzić pewną modyfikację [8]. Jeżeli wartością odcinaną okaże się ta, która pierwotnie zawierała brak, losową metodę „hot deck” należy zastosować tak jak to zostało opisane powyżej. Jeżeli jednak wartością odcinaną będzie wartość zaobserwowana, każda z wartości wylosowanych powinna być skorygowana – dodana do niej powinna zostać wielkość $\hat{\theta}_{r-1,i} - \hat{\theta}_r$, obliczona na podstawie wartości zaobserwowanych, gdzie r oznacza liczbę wartości zaobserwowanych.

W takim sposobie szacowania wariancji statystyki zostaje uwzględniona jej dodatkowa zmienność wynikająca z niekompletności danych, dzięki czemu możemy uniknąć niedoszacowania tej wariancji.

4. Metoda bootstrap

Metoda bootstrap została przedstawiona przez Bradleya Efrona w roku 1979 [1] jako pewna modyfikacja metody jackknife. Nazwa wybrana dla tej metody

związana jest z wyrażeniem „pull oneself up by bootstraps”, które oznacza „wydobyć się z opresji za pomocą własnych sił”. Metoda ta podobnie jak jackknife opiera się na charakterystycznym tylko dla niej rodzaju próby.

W celu otrzymania próby typu bootstrap X_b^* dokonujemy n -krotnego losowania ze zwracaniem spośród elementów oryginalnej próby X . Próba bootstrap składa się z elementów próby oryginalnej, z których każdy może występować raz, dwa razy itd., lub nie występować w niej wcale. Metoda bootstrap wymaga wygenerowania B takich prób [2].

Następnym krokiem jest obliczenie replik bootstrapowych postaci $\hat{\theta}_b^* = T(X_b^*)$, gdzie X_b^* oznacza próbę typu bootstrap X_b^* . Estymatorem typu bootstrap wariancji estymatora $\hat{\theta}_n$ nazywamy wielkość:

$$\hat{v}_B = \frac{1}{B-1} \sum_{b=1}^B (\hat{\theta}_b^* - \hat{\theta}_{(\cdot)}^*)^2, \quad \text{gdzie} \quad \hat{\theta}_{(\cdot)}^* = \frac{1}{B} \sum_{b=1}^B \hat{\theta}_b^*.$$

Estymator obciążenia estymatora $\hat{\theta}_n$ parametru θ ma postać:

$$\text{bias}_B = \hat{\theta}_{(\cdot)}^* - \hat{\theta}_n.$$

Aby otrzymać dobrą estymację wariancji estymatora $\hat{\theta}_n$, wystarczy często nawet niewielka liczba replik (20-200). Zastosowanie metody bootstrap nie kończy się jednak na estymacji błędu standardowego estymatora. Za jej pomocą możliwa jest zarówno estymacja punktowa, jak i przedziałowa, testowanie hipotez oraz wiele innych analiz. W każdym jednak przypadku potrzebna jest inna liczba replik (np. gdy do estymacji błędu standardowego wystarczy wygenerowanie 50 replik, to w przypadku estymacji przedziałowej potrzeba ich od 1000 do 10 000) [2].

Podstawę użycia estymatora typu bootstrap wariancji stanowi dowodzona jego mocna zgodność. Dowód ten znajduje się w pozycji [9].

4.1. Metoda bootstrap w analizie danych niepełnych

W analizie danych niepełnych proces imputacji podobnie jak dla metody jackknife powtarzany jest dla każdej próby typu bootstrap oddzielnie. Na początku generowane są próby X_b^* . Następnie w każdej z nich dokonujemy imputacji przy użyciu wybranej metody i obliczamy repliki $\hat{\theta}_b^*$, na których podstawie estymujemy wartość oraz wariancję interesującej nas statystyki [9].

5. Podsumowanie

Metody klasyczne bazują na założeniach dotyczących rozkładu cechy, o którym wnioskuje się dzięki przeprowadzanym testom na określoną postać rozkładu.

Następnie przy założeniach, że próba ma rozkład bliski znanemu rozkładowi przeprowadzane jest wnioskowanie. W przypadku metod bootstrap i jackknife nie musimy wnioskować o rozkładzie cechy. Kwestie dyskusyjne, jakie mogą się tu pojawić, to wielkość i rodzaj próby, przy której metody te mogą być stosowane, liczba generowanych prób typu bootstrap oraz to, w jakich sytuacjach stosować jedną lub drugą metodę.

W niektórych przypadkach dzięki metodzie jackknife możemy otrzymać dobre wyniki dużo mniejszym kosztem w porównaniu z metodą bootstrap. Na przykład bootstrap wymaga wygenerowania 200 replik, podczas gdy w jackknife potrzeba ich tylko n (co oczywiście ma sens tylko, gdy $n < 200$). Jednak pole działania jest tu ograniczone, ponieważ metoda jackknife zawodzi, gdy funkcja $T(X_n)$ nie jest wystarczająco gładka. Z tego względu nie uda się uzyskać dobrych oszacowań np. dla mediany. Metoda jackknife zazwyczaj jest efektywna pod warunkiem, że wartość statystyki nie zmienia się drastycznie wraz z małymi zmianami danych. Można pokazać, że w przypadku mediany estymator wariancji typu jackknife nie jest zgodny, podczas gdy estymator typu bootstrap jest [6].

Z drugiej strony wydawać się może, że skoro w metodzie jackknife jesteśmy w stanie wygenerować jedynie n prób typu jackknife, to metoda ta jest mniej efektywna niż np. metoda bootstrap. Okazuje się, że dla wystarczająco gładkich statystyk estymatory typu jackknife mogą być postrzegane jako estymatory typu bootstrap. Dzięki temu związkowi w szczególnych przypadkach jackknife i bootstrap dają prawie identyczne rezultaty. Wystarczy wtedy stosować metodę jackknife. Ma to miejsce np. w przypadku statystyk liniowych, czyli postaci:

$$\hat{\theta} = \mu + \frac{1}{n} \sum_{i=1}^n \alpha(x_i), \text{ gdzie } \mu = \text{const} [9].$$

Literatura

- [1] Efron B. (1979), *Bootstrap Methods: Another Look at the Jackknife*, „Ann. Statist.” 7, 1-26.
- [2] Efron B. (1993), *An Introduction to the Bootstrap*, Chapman & Hall, London.
- [3] Jermakov S.M. (1976), *Metody Monte Carlo i zagadnienia pokrewne*, PWN, Warszawa.
- [4] Kish L., Frankel M.R. (1973), *Inference from Complex Samples*, „Journal of the Royal Statistical Society”, Series B, 36.
- [5] Koronacki J., Mielniczuk J. (2001), *Statystyka dla studentów kierunków technicznych i przyrodniczych*, WNT, Warszawa.
- [6] Politis D.N., Romano J.P. (1999), *Subsampling*, Springer-Verlag.
- [7] Quenouille M. (1949), *Approximate Tests of Correlation in Time Series*, „Journal of the Royal Statistical Society”, B 11, 18-84.
- [8] Rao J.N.K., Shao J. (1972), *Jackknife Variance Estimation with Survey Data under hot Deck Imputation*, „Biometrika”, 79.
- [9] Shao J., Tu D. (1995), *The Jackknife and Bootstrap*, Springer-Verlag, New York.

THREE NONCLASSICAL METHODS OF STATISTICAL INFERENCE: MONTE CARLO, BOOTSTRAP AND JACKKNIFE

Summary

This paper describes and compares three methods of statistical inference: Monte Carlo, bootstrap and jackknife. Monte Carlo methods are presented as a basis for two other methods. The wide range of applications areas of jackknife and bootstrap is shown only in selected aspects. Their application to estimating parameters and assessing statistical accuracy is considered. Moreover bootstrap and jackknife is presented as data imputation techniques.